Development of a Mexican Spanish Synthetic Voice Using Synthesizer Modules of Festival Speech and HTS-Straight

Abel Herrera Camacho and Fernando Del Rio Ávila

***Abstract*—A new Mexican Spanish voice was created using 4 festival modules (Clunits, Clustergen,Multisyn and HTS), as well as an additional database was created with Straight processing. All voices were created using the same database to allow for consistency and for easier comparison of the output. Once these voices have been created they can be used as a baseline for further development in Mexican Spanish speech synthesis. Except for the Multisyn module, which has some problems due to coverage, results are acceptable, with the HMM designed voices having the best quality.**

***Index Terms*—Speech synthesis, straight, festival.**

I. Introduction

The new speech synthesis techniques: hts and straight, have improved the recent research on this field. Specifically, looking for natural speech the HTS-Straight technique has had excellent results [1], [2]. The HTS technique has provided buzzy speech for American English Language [3]. We tried to check these results for Mexican Spanish Language.

However, the HTS technique is easy to insert in the well know Festival System, with some modification to the system to account for missing parameters, so it is easy to compare HTS with traditional modules of Festival as Clunits and Clustergen.

In this article, we compared the quality of synthesized speech of Festival modules Clunits, Clustergen,Multisyn and HTS for Mexican Spanish.

The use of these modules and techniques requires not only a straightforward adaption to Mexican Spanish. In some cases, especially for the HTS-Straight Technique, the adaption required a strong software redesign.

Other purpose of the article was to check the buzzy effect of the HTS Module for Mexican Spanish, more carefully than we have done before [4].

II. Basic Description of Synthetizer Modules

A. Clunits[5]and Multisyn[6]Modules

The Clunits (Cluster Units) method works by extracting a list of phones from a set of prerecorded phrases, including their prosodic context, generating a CART (Classification and Regression Tree), that according to context will give a set of possible segments to use depending on context at synthesis time.

During synthesis, a target is generated for each phone to be synthesized (based on its context). Once a set of possible units is extracted from the CART for each target, an optimal path is generated, concatenating the units that will generate the smallest weight throughout the phrase.

Multisyn works in a similar fashion, replacing phone selection with diphones.

B. HTS [7] and Clustergen [8] Modules

The HTS module is based on Hidden Markov Models. These HMMs are used to generate the decision trees used to select the optimal set of parameters during synthesis time.

The HTS module uses three sets of parameters (Mel Cepstral, F0 and phone duration). Each set of parameters is extracted from the database independently from the others, allowing for prosody modifications at the cost of some distortion due to the source/filter model used during synthesis time.

Clustergen works the same as HTS, with only some modifications on the way the HMMs are generated.

C. Straight [9]

Straight synthesis is done using HTS, by replacing the Mel-Cepstral parameters with Straight parameters. However, the HTS module in festival does not allow for the use of Straight parameters.

In this case, Festival is used exclusively for prosodic analysis, using this output to feed a set of external applications (HTK, HTS and Straight (using MATLAB) for parameter selection and synthesis.

III. Database Recording and Labeling

The database used for the voice generation consists of approximately 60 minutes of poetry. Each phrase (around 1000) was stored in a separate file. A transcription file was also generated to be used for the phone labeling of each audio file.

Additionally, a set of Festival SCHEMA files were created. These files contain the rules for extracting the phones from the text input, including some exceptions and handling of numbers, dates and basic formatting.

The Schema files were adapted from 2 existing Castilian voices (el_diphone, provided with festival, and the voices created for the Guadalinex Project [10]), using a modified phoneset to better model the Mexican variant of Spanish, and we adopted small changes not used before [4].

For this database 27 phones were used (17 consonants and 10 vowels (Spanish has only 5 vowel sounds, but stressed

Manuscript received October 25, 2012; revised November 29, 2012.

The authors are with the Universidad NacionalAutonoma de Mexico (UNAM) (e-mail: abelhc@hotmail.com).

and un-stressed variants are used as different phones for easier processing), see Table I.

Each of the consonant phones is classified according to three categories, while the vowels were classified in four categories.

In the case of all HMM based voices, the labeling was carried out with the EHMM labeler, using the same labeling in all cases. The labeler outputs the starting, middle and ending point of each recorded phone. In all cases some manual verification and correction was carried out.

A. Consonant Classification

By Type:

1) S (occlusive): The oral and nasal cavities are closed, there is no air flow.
2) F (fricative): There is constant friction in the articulation point, the air flow is not completely restricted.
3) A (africative): Formed by an occlusive sound, followed by a fricative.
4) N (nasal): The oral cavity is closed, air flows through the nose.

By Articulation:

1) B (bilabial): Lips against each other
2) L (labiodental): Lower lip against teeth.
3) D (dental): Tongue against teeth
4) A (alveolar): Tongue against the base of the teeth
5) P(palatar): Tongue against the hard palate
6) V(velar): Tongue against the soft palate.

By Voicing:

1) Voiced (+): Vocal cords vibrate
2) Unvoiced (-): No vibration

B. Vowel Classification

By Height: the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw (Low, Mid or High).

By Position: the position of the tongue during the articulation of a vowel relative to the back of the mouth (Front, Mid or Back).

Lip Rounding: The lips are (+) or not (-) in a rounded position.

Stressed: The vowel is (+) or not (-) stressed.

IV. Training Process

A. Festival Training

The three festival based voices (Clustergen,Multisyn and Clunits) were trained using the Festvox software.

Clunits [5]: For clunits, extracts all instances of each phone and clusters them according to their context. During synthesis time, the context is extracted and the corresponding cluster is extracted, selecting the set of segments with smaller cost based on their Cepstral parameters.

Multisyn [6]: For multisyn, diphones are used, selecting an optimal path at synthesis time. If a diphone can't be found, a backoff module is used to replace with an appropriate replacement.

Clustergen [8]: Clustergen training is HMM based, creating a set of CART trees (MelCepstral, F0 and duration), each set of parameters calculated independently from each other. Clustergen is phone based, clustering phones according to their context.

TABLE 1: Phones Used in the Voicebank

	Consonants			
Name	Vowel	Type (*)	Articulation (**)	Voicing
p	-	s	L	-
t	-	s	D	-
k	-	s	V	-
b	-	s	L	+
d	-	s	D	+
g	-	s	V	+
f	-	f	B	-
s	-	f	A	-
x	-	f	V	-
ch	-	a	P	-
m	-	n	l	+
n	-	n	a	+
ny	-	n	p	+
l	-	l	a	+
ll	-	l	p	+
r	-	l	a	+
rr	-	l	a	+

	Vowels				
Name	Vowel	Height	Stressed	Position	Lip Rounding
(sil)	-	-	-	-	-
A	+	Low	-	Mid	-
E	+	Mid	-	Front	-
I	+	High	-	Front	-
O	+	Mid	-	Back	+
U	+	High	-	Back	+
a1	+	Low	+	Mid	-
e1	+	Mid	+	Front	-
i1	+	High	+	Front	-
o1	+	Mid	+	Back	+
u1	+	High	+	Back	+

B. HTS Training.

For HTS training while labeling was done using the EHMM tool provided with festival, training itself is used the HTK tools.

For HTK training, a set of ‘questions’ must be provided, that will contain the context information that will be used for the creation of the decision tree. This information must match the context information generated by the Festival labeling, but manual adjustments can be done.

The HTK/HTS tools also provide a set of parameter that allows easy modification of the parameterization of the audio data. Due to this fact, different HTS voices were created to validate the effect of different factors in the quality of the synthetic voice.

C. HTS Parameters

Frequency Warping: This parameter allows the use of Cepstral or Mel-Cepstral parameters. Voices were created using no frequency warping and Mel Scale. As we expected,

better results were obtained when using the Mel warped parameters.

Gain: Log or Linear gain. Notice that the festival HTS module must be modified for log gain to work from festival. Log gain provides slightly better results, but at the cost of a much higher training time.

Gamma: This parameter affects the reconstruction filter parameters placement of poles and zeroes. Values of -1,-1/3 and 0 were used. Best results were obtained using -1/3, but care must be taken with the number of cepstralparameters, as the filter can become unstable.

Number of Cepstral Coefficients: Vectors consisting of 12, 24 and 36 parameters were used. With 12 coefficients the reconstructed signal is too distorted. With more coefficients the reconstructed signal is clearer, but with a high number of coefficients the filter can become unstable, see Fig. 1 and 2.

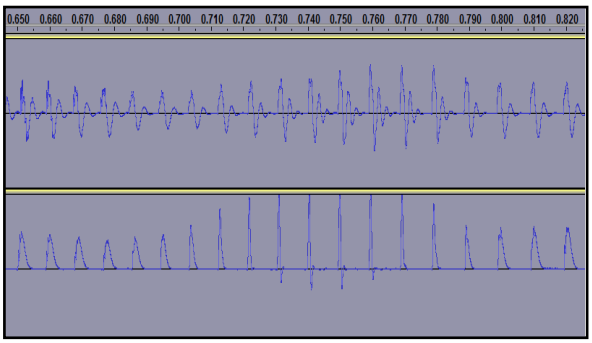

Fig. 1. Signal with 36 and 12 coefficients.

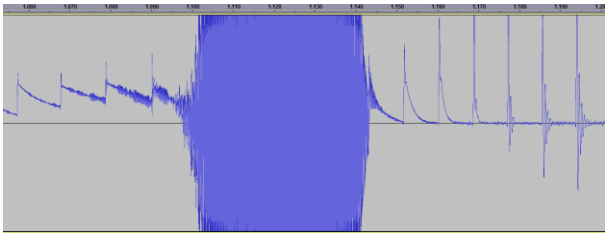

Fig. 2. Signal with 36 coefficients and Gamma=-1/3.

Number of states per phone: Each phone is divided into a number of HMM states, see Fig. 3. Voices were created with 3,5 and 7 states.

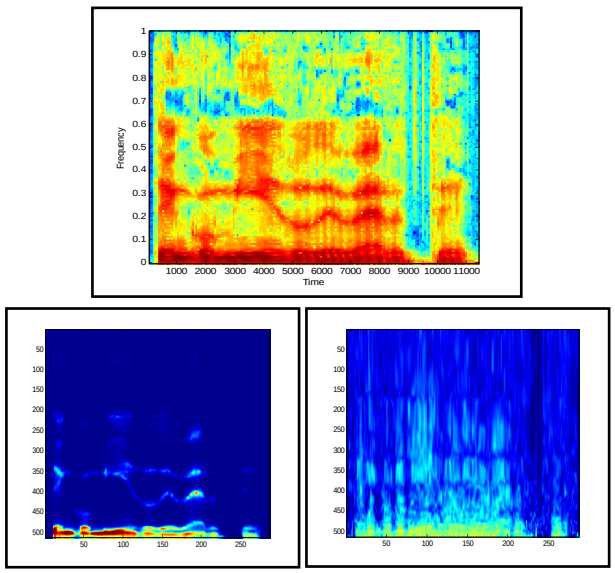

Fig. 3. a) Original spectrum. b) Smoothed spectrum. c) Aperiodicity data

D. HTS-Straight Training

In the case of HTS-Straight, the training is done in a similar fashion to standard HTS training, but replacing the Mel-Cepstral parameters with straight parameters.

Straight processing divides the signal into three discrete set of parameters:

1) A pitch track, with the fundamental frequency of the speech signal
2) A smoothed spectrogram, which contains the periodic, slowly evolving part of the spectrum
3) An aperiodicity signal, containing the residual from the smoothed spectrum.

The relationship between the original spectrogram and the STRAIGHT representation is shown in Fig. 3.

V. Results

The voices synthesized with the concatenative approach were found to have discontinuities at synthesis time. Of particular notice is the Multisyn module, as the database used was not phonetically balanced, resulting in high discontinuities and gaps in the synthesized speech. As the other modules are phone-based, they avoid this problem.

The voices synthesized by HTS and HTS-Straight were valuated as very natural by four linguistic experts in our lab, more the second one. We did not used a MOS test, we preferred experts to check fundamental features of the voice.

The buzzy effect is not relevant for Mexican Spanish, three of four experts did not heard it, and the only one heard "something unnatural" in the hts technique but not defined it as buzzy voice.

We designed an interface for easy use of these techniques, and promptly will be ready for free internet access.

From 3 to 5 states per phone there is marked improvement. Beyond 5 states, this is minimal. And the experts split decisions about quality rise of the voice from 5 to 7 states.

Finally, we hope these experiences will help researchers for Spanish Language in the use of these techniques.

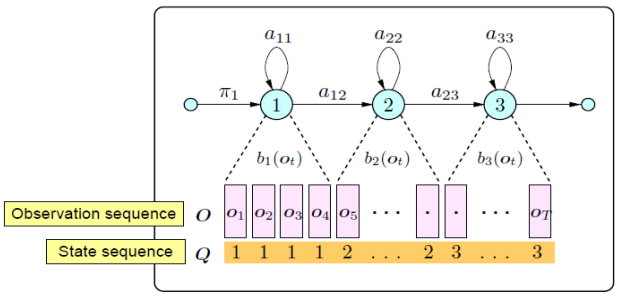

Fig. 4. Phone subdivision into HMM states. (From [11])

References

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of IEEE ICASSP*, 2000, pp. 1315–1318.

[2] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. of IEEE ICASSP*, 1997, pp. 1303–1306.

[3] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in*Proc. of 2002 IEEE SSW*, September, pp. 227-230, 2002.

[4] D. Fernando and H. Abel, "Development of a Mexican Spanish HMM-Based Synthetic Voice," presented at the *18th Mexican*

International Congress on Acoustics, Cholula, Mexico 16-18 November 2011.
[5] W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. [Online]. Available: https://www.era.lib.ed.ac.uk/bitstream/1842/1236/1/Black_1997_b.pdf
[6] K. Richmond and S. King. Multisyn: open-domain unit selection for the festival speech synthesis system. Robert A.J. Clark. [Online]. Available: http://peer.ccsd.cnrs.fr /docs/00/49/91/77/PDF/PEER_stage2_10.1016%252Fj.specom.2007.01.014.pdf
[7] H. Zen, T. Nose *et al*.,The HMM-based speech synthesis system (HTS) version 2.0. [Online]. Available: http://mir.cs..nthu.edu.tw/users/heycat/relatedPapers/2_The%20HMM-based%20Speech%20Synthesis%20System%20Version%202.0.pdf
[8] A. W Black. CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. [Online]. Available: http://www-2.cs.cmu.edu/~awb/papers/is2006 /IS06139 4.PDF
[9] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1303 –1306, vol. 2, 1997.
[10] Desarrollovoces sinteticasguadalinex. [Online]. Available: http://forja.guadalinex.org/frs/download.php/155/DesarrolloVocesSinteticasGuadalinex.pdf
[11] F. Itakura, "Fundamentals of speech analysis and synthesis and its application to speech coding," *IEI CE FM06-2-1*, July 2006.

Abel Herrera C. graduated in 1979 as a mecanical-electrical engineer, aquiring the Master in Electrical Engenieering in 1985 and his Ph.D. in 2001, all at the University of Mexico (UNAM), the Ph.D. with help from the University of California-Davis. He carried out a post-doctoral internship at Carnegie Mellon University in 2001, as well as a research internship at the University of Southern California. Author of over 50 papers in speech codification, recognition and synthesis. Head of the Voice Processing lab, has taught at the University since 1979.

Fernando Del Rio graduated in 2003 as a computer engineer at the University of Mexico (UNAM), obtained his master in electrical engineering in 2005. He also has a master in engineering manager from McNeese State University.